同行者コンテキスト依存の文書抽出およびトピック解析

深澤佑介　太田順

東京大学　人工物工学研究センター

yusuke.fukazawa@gmail.com

概要　本稿では、著者らが提案した同行者依存のトピックモデルの評価を行う。質的評価として同行者のトピックに含まれる単語を確認し、妥当なモデル化が行われていることを確認した。量的評価としてKL-Divergenceにより提案手法、LDAの双方のトピック間の分離性能を確認し、提案手法が優れていることを確認した。

キーワード　同行者、コンテキスト、トピックモデル

1 はじめに

ユーザのコンテキストはユーザが生成する文書のトピックに大きく影響を与える重要な要素である。Adomaviciousらは、コンテキストとは,「時間」「場所」「同行者」であると定義している[1]。コンテキストを考慮したトピックモデルとして「時間」[2][3]「場所」[4][5]に応じたトピックモデルが数多く提案されている。しかしながら、時間と場所が同一でも同行者の有無、同行者が誰かによってトピックは変わる。深澤らは、「同行者」をトピックを変化させるパラメータとして考慮したトピックモデルを提案している[6]。提案モデルは、同行者依存の単語を抽出するため、各単語ごとに、次の3点：1)同行者によって決まる単語、2)ユーザの興味によって決まる単語、3)ユーザの興味や同行者に無関係に決まる単語を分類するスイッチ変数を導入している点が特徴である。本稿では、「同行者」をトピックを変化させるパラメータとして考慮したトピックモデルと過去のLDA[7]によるモデル（比較手法）の質的評価および量的な評価を行う。

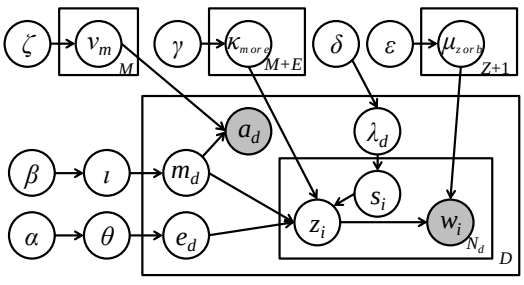

図 1　提案モデル（m_dは文書dの同行者潜在クラス、a_dは文書dの同行者、e_dは文書の潜在嗜好クラス、zは文書のトピック、s_iはスイッチ変数を表す。その他のパラメータは[6]を参照のこと。）

2 同行者依存のトピック抽出

一般に明示的に同行者に関する情報が付与された文書情報はない。ここでは、文書情報から同行者に関する情報を抽出することで学習データを構築する。本稿では、Twitterにおいて同行者を含む投稿文を評価用データの対象とする。ここでは、「with 同行者」となっている投稿を抽出する。データ抽出には検索エンジンBingを用いた。なお対象となる同行者は英語の学習サイト(http://usefulenglish.ru/vocabulary/jobs-professions-occupations)から抽出した。同行者のデータセットはプライベートにおける同行者を対象とし、「with 同行者」の形式を投稿文中に含むTwitterの投稿を10件づつ抽出した。同行者の総数は計79個である。総単語数は計617個である。また、図2の各パラメータの初期値は[6]に従う。

3 質的評価

提案モデルおよびLDAによる学習結果をそれぞれ、表1および表2に示す。表では、各潜在同行者クラスと、それに対応する潜在トピッククラスの対応関係を示している。各潜在同行者クラスにはそのクラスに属する同行者の集合、および単語が記載されている。各潜在同行者クラスの単語は、各潜在同行者クラスで最も属する確率の高い潜在トピッククラスを抽出、その潜在トピッククラスに属する単語を表示している。表1からわかるように、潜在同行者クラスの同行者（例：bride、 fiance）と対応する潜在トピッククラスの単語を確認すると、engaged、groomなど同行者に関係のある単語が上位の単語として抽出されている。一方、表2に潜在同行者クラスの同行者（例：wife、 mistress）と対応する潜在トピッククラスの単語を確認すると、divorce、thanksなど関係のある単語もあるが上位の単語は

blog、join など一般的な単語が占められている。これは、提案モデルにおいてはスイッチ変数により同行者に特有の単語を絞っているが LDA ではそのような機構がないため、同行者に依存しない一般的な単語も混在していると考えられる。以上より、提案モデルによって妥当な分類結果が得られていることが分かる。

同行者クラスに属する同行者について考察する。提案手法では、(girlfriend、mistress)、(bride、fiance)、(child、 daughter) など関連のある同行者が同一の同行者クラスに属している。一方、LDA によるモデルのほうでは (bride、child)、(child、my relatives)、(husband、brother-in-law) など関係が少ないと思われる同行者が同一の同行者クラスに属している。これは、LDA では、同行者に依存する単語／非依存の単語が混在したトピックに基づき同行者を分類しているため、適切に同行者を分類することができなかったと考えられる。

一方、提案手法においても john、peter などの人名が分類されているが、人名はどのような同行者にもなりえるため同行者を予測する際のノイズとなる可能性がある。例えば、john は father にも son にもなりえる。このような単語が上位の単語として分類されてしまった原因として、モデルを学習する際に学習データのデータ量が少なく特定の文書の影響を受けているためと考えられる。今後、学習データ量を増加し大規模なデータを用いて学習することで精度を向上させる。

表 1　提案モデルによる同行者クラスの単語分布

Class #	1	2	3	4	5
associated companion	girlfriend	bride	child	husband	parents
	mistress	fiance	daughter	my family	grandmother
associated words	night	favorite	support	single	home
	divorce	engaged	separated	god	visit
	tips	cool	photos	hate	following
	club	groom	really	business	date
	feelings	cooking	rtfqp	john	learn
	lovely	set	grandchild	school	long
	food	meet	read	guy	waiting
	sxsw	peter	funny	wine	office

表 2　LDA による同行者クラスの単語分布

Class #	1	2	3	4	5
associated companion	wife	bride	child	husband	sister
	mistress	child	my relatives	brother-in-law	parents
associated words	blog	help	free	man	tonight
	join	need	rich	ask	news
	husband	man	people	trip	living
	hour	home	son	anne	right
	divorce	grandparents	dads	family	button
	instantly	bride	better	account	thanks
	online	button	family	parents	updated
	thanks	folks	grandmother	engaged	samantha

4 量的評価

一般にトピックモデルにおいて生成されたトピック同士は意味的により分離していることが望ましい。トピック間の分離性能を量的に評価するため、KL-Divergence により評価を行った。クラス z_1 とクラス z_2 間の KL-Divergence は次式で表される。

$$KL(z_1, z_2) = \sum_{w} p(w \mid z_1) \log \frac{p(w \mid z_1)}{p(w \mid z_2)}$$

ここで、KL-Divergence が大きいほどその二つのトピックは互いに分離性が高いといえる。一方、0 の場合は、二つのトピックは全く同一である。クラス数 K=10 で実施した際の結果を図 2 に掲載する。図 2 からわかるように、提案手法のほうがよりトピック間の分離性が高くなっている。これは、提案モデルにおいてはスイッチ変数により同行者に特有の単語を絞っているが LDA ではそのような機構がなくトピック間で共通の単語が混在しているためと考えられる。

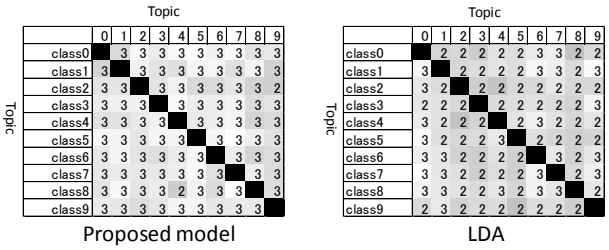

図 2　KL-Divergence による評価

5 結論

本稿では、同行者依存のトピックの発見モデルを評価した。今後は、その他のコンテキスト（時間や位置）も考慮したトピックモデルのモデル化を目指す。

参考文献

[1] G. Adomavicius and A. Tuzhilin, "Context-Aware Recommender Systems," *Recommender Systems Handbook*, pp.217-253, 2011.

[2] X. Wang and A. McCallum, "Topics over time: a non-markov continuous-time model of topical trends," *Proc. of KDD*, pages 424-433, 2006.

[3] N. Kawamae, "Trend analysis model: trend consists of temporal words, topics, and timestamps," *Proc. of WSDM*, 317-326, 2011.

[4] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, "A latent variable model for geographic lexical variation," *Proc. of EMNLP*, pp. 1277–1287, 2010.

[5] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, K. Tsioutsiouliklis, "Discovering geographical topics in the twitter stream," *Proc. of WWW*, 769-778, 2012.

[6] 深澤 佑介, 太田 順, 同行者依存のトピック発見モデル, 情報処理学会研究報告, 2012-MBL-63, 3, 1/9, 2012.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation, *Journal of Machine Learning Research*, pp. 993-1022, 2003.